Segmentação de Clientes com Machine Learning: Clusterização e Estratégias de Personalização Digital para Impulsionar Vendas

William N. Filho

Resumo. *Este relatório se propõe a aplicar técnicas de mineração de dados em dados de compras em supermercado para segmentação de clientes e identificação de regras de associação entre categorias de produtos por meio da linguagem Python, com o objetivo de compreender perfis de consumo e propor estratégias para impulsionar as vendas e aprimorar a experiência do usuário. As soluções propostas envolvem sistemas de recomendação, ordenação personalizada de itens e campanhas de marketing segmentadas. A metodologia envolve uma análise exploratória de dados (EDA), um estudo comparativo entre os algoritmos de clusterização K-means e Gaussian Mixture Model (GMM) baseado em métricas como Silhouette Score e Davies-Bouldin Index, a definição do número adequado de clusters e a aplicação do algoritmo Apriori para extrair padrões de compra, permitindo a formulação de estratégias de marketing eficazes e personalizadas.*

1. Objetivo

Este relatório tem como objetivo segmentar clientes de um supermercado com base em seus comportamentos de compra e identificar padrões de associação entre categorias de produtos, visando compreender diferentes perfis de consumo e propor estratégias para impulsionar as vendas e aprimorar a experiência do usuário. As estratégias incluem a implementação de sistemas de recomendação personalizados, ordenação estratégica de produtos e de campanhas de marketing direcionadas por perfil de cliente. Para alcançar o objetivo do estudo, foram aplicadas técnicas de mineração de dados e aprendizado de máquina, como análise exploratória, algoritmos de clusterização e extração de regras de associação, utilizando a linguagem Python.

2. Metodologia

Este estudo foi conduzido em quatro etapas principais, utilizando a linguagem Python e as bibliotecas pandas, numpy, openpyxl, scikit-learn, mlxtend, joblib, matplotlib, seaborn e plotly.

1. **Análise Exploratória de Dados (EDA)**: Realizou-se uma análise detalhada dos dados, dividida em quatro tópicos: (i) padrões temporais dos pedidos, (ii) identificação de produtos, seções e departamentos mais populares, (iii) comportamento de compra dos clientes, e (iv) análise de recompras.
2. **Clusterização de Clientes**: Construiu-se uma matriz de proporções representando a distribuição relativa de categorias de produtos compradas por cada cliente. Para a segmentação, comparou-se os algoritmos K-means e Gaussian Mixture Model (GMM) com base nas métricas Silhouette Score e Davies-Bouldin Index. O K-means foi selecionado por seu melhor desempenho, e o número ideal de clusters foi determinado com apoio do Elbow Method, priorizando uma segmentação eficaz e interpretabilidade dos perfis de clientes. Após a clusterização, os perfis

de cada grupo foram analisados com base na média proporcional de consumo por departamento.

3. **Regras de Associação**: Aplicou-se o algoritmo Apriori para extrair regras de associação entre categorias de produtos compradas com frequência dentro de cada cluster. As regras mais fortes foram selecionadas com base em valores elevados de lift e confidence.
4. **Aplicação dos Resultados**: Com base nas regras extraídas, foram propostas estratégias para personalização do ambiente digital, incluindo sistemas de recomendação e ordenação estratégica de produtos exibidos ao usuário, segmentados por grupo de cliente. Além disso, foram propostas campanhas de marketing direcionadas com base no perfil de cada grupo. Essas ações visam impulsionar as vendas, aprimorar a experiência do usuário e aumentar a retenção de clientes.

3. Apresentação da Base de Dados

A base de dados utilizada neste estudo é adaptada de um repositório disponível no Kaggle (https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis), com a finalidade de simular transações de compras online em um supermercado. Ela reúne informações detalhadas sobre o histórico de compras dos usuários, incluindo os produtos adquiridos, a recorrência das recompras, a categorização de produtos por departamentos e seções, além da sequência em que os pedidos foram realizados por cada cliente. A base utilizada pode ser acessada em https://github.com/WNabhan/customer-segmentation-analysis.

Os metadados de cada arquivo da base estão representados nas tabelas a seguir.

Tabela 1. Metadados de aisle.csv

Campo	Tipo	Descrição
aisle_id	Numérico	Identificador da seção
aisle	Categórico	Nome da seção

Tabela 2. Metadados de department.csv

Campo	Tipo	Descrição
department_id	Numérico	Identificador do departamento
department	Categórico	Nome do departamento

Tabela 3. Metadados de products.csv

Campo	Tipo	Descrição
product_id	Numérico	Identificador do produto
product_name	Categórico	Nome do produto
aisle_id	Numérico	Identificador da seção
department_id	Numérico	Identificador do departamento

Tabela 4. Metadados de orders.parquet

Campo	Tipo	Descrição
order_id	Numérico	Identificador do pedido
user_id	Numérico	Identificador do usuário
eval_set	Categórico	Indica se o pedido pertence ao conjunto prior, train ou test
order_number	Numérico	Número do pedido do usuário
order_dow	Numérico	Dia da semana do pedido do usuário 0 (domingo) a 6 (sábado)
order_hour_of_day	Numérico	Hora do pedido do usuário 0 (domingo) a 6 (sábado)
days_since_prior_order	Numérico	Intervalo de dias entre o pedido atual e o pedido anterior

Tabela 5. Metadados de order_products.parquet

Campo	Tipo	Descrição
order_id	Numérico	Identificador do pedido
product_id	Numérico	Identificador do produto
add_to_cart_order	Numérico	Ordem de adição ao carrinho do produto
reordered	Numérico	Indica se o produto é uma recompra em relação ao histórico do usuário (0 = primeira compra, 1 = item já adquirido anteriormente)

4. Preparação dos Dados

Para realizar a análise exploratória dos dados, segmentar os usuários e extrair as regras de associação, foi feita uma mesclagem dos arquivos apresentados anteriormente, tendo como resultado um conjunto de dados único, possuindo 4.986.345 instâncias.

	user_id	order_id	order_number	order_dow	order_hour_of_day	days_since_prior_order	product_name	aisle	department	add_to_cart_order	reordered
0	1	1	1	3	12	NaN	Organic Raw Agave Nectar	honeys syrups nectars	pantry	1	0
1	1	1	1	3	12	NaN	Organic Soba	asian foods	international	2	0
2	1	1	1	3	12	NaN	Organic Red Cabbage	fresh vegetables	produce	3	0
3	1	1	1	3	12	NaN	Organic Shredded Carrots	packaged vegetables fruits	produce	4	0
4	1	1	1	3	12	NaN	Organic Red Onion	fresh vegetables	produce	5	0
5	1	1	1	3	12	NaN	Red Raspberries	packaged vegetables fruits	produce	6	0
6	1	1	1	3	12	NaN	Organic Cilantro	fresh herbs	produce	7	0
7	1	1	1	3	12	NaN	Organic Blackberries	fresh fruits	produce	8	0
8	1	1	1	3	12	NaN	Whole Vitamin D Milk	milk	dairy eggs	9	0
9	1	1	1	3	12	NaN	Jicama	fresh vegetables	produce	10	0

Figura 1. Recorte das 10 primeiras instâncias da base de dados consolidada

4.1 Checagem dos dados

Antes de iniciar a EDA, foi feita uma checagem da integridade e consistência dos dados. Para isso, foram analisados e avaliados os valores únicos de entrada de cada campo, a presença de valores nulos e a ocorrência de valores diferentes de ’user_id’, ‘order_number', 'order_dow', 'order_hour_of_day', 'days_since_prior_order' para um mesmo ‘order_id’.

Ocorreram valores nulos apenas na coluna 'days_since_prior_order', porém foi averiguado que todas as ocorrências pertencem às instâncias cujo valor de ‘order_number’ é 1. Além disso, não foram constatados valores conflitantes de ’user_id’, 'order_number', 'order_dow', 'order_hour_of_day', 'days_since_prior_order' para um mesmo ‘order_id’ e as quantidades de valores únicos de cada campo foram: ‘user_id’: 30000; ‘order_id’: 492817; ‘order_number’: 100; ‘order_dow’: 7; ‘order_hour_of_day’: 24; ‘days_since_prior_order’: 31; ‘product_name’': 44439; ‘aisle’: 134; ‘department’: 21; ‘add_to_cart_order’: 93 e ‘reordered’: 2.

5. Análise Exploratória dos Dados

A EDA foi dividida em 4 tópicos e foi realizada em Python. Para isso foram utilizadas as bibliotecas Pandas (para manupilação e consolidação), Matplotlib, Seaborn e Plotly (para plotagens e visualização).

5.1 Análise de Padrões Temporais dos Pedidos

Para entender os padrões de compra temporais, os dados foram consolidados por dias da semana e hora do dia. Para facilitar a interpretação foi criada uma coluna ‘dia_da_semana’ com valores de segunda a domingo, tomando como base a coluna ‘order_dow’.

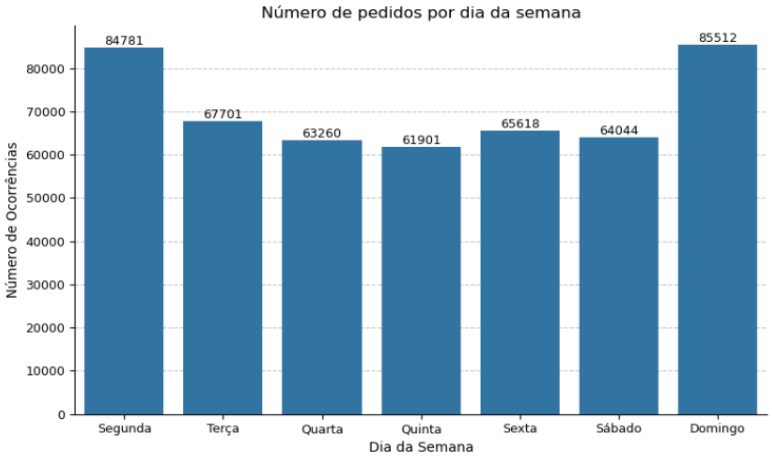

Figura 2. Gráfico de contagem de pedidos por dia da semana

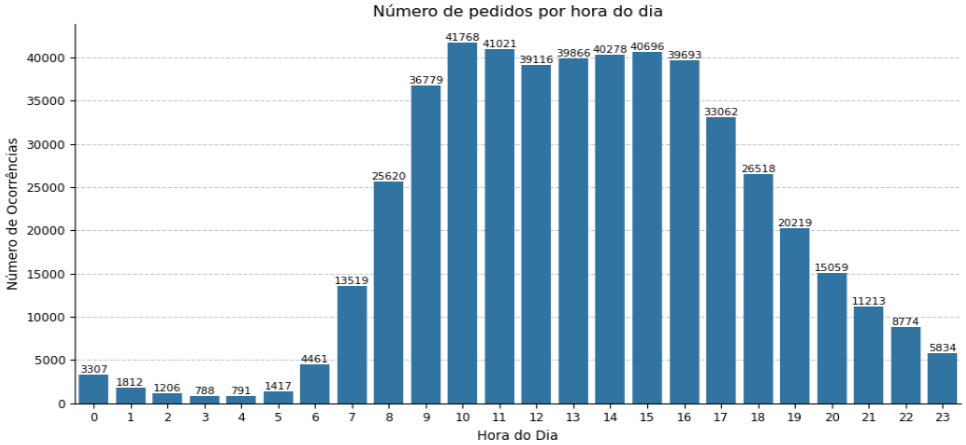

Figura 3. Gráfico de contagem de pedidos por hora do dia

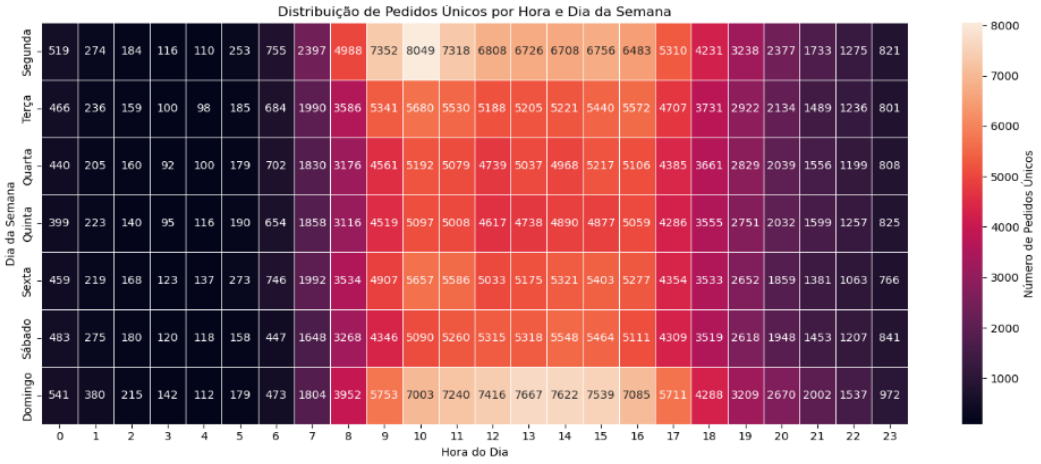

Figura 4. Heatmap de pedidos por dia da semana e hora do dia

Uma forma estratégica de utilizar esses gráficos em um cenário de compras presenciais em supermercados é filtrando os dados por departamento, seção ou produto para identificar padrões de compra específicos ao longo dos dias da semana e horários do dia, permitindo compreender os momentos de maior demanda e, assim, otimizar a alocação

de equipes para reposição de estoque e atendimento em caixa, garantindo maior eficiência operacional e melhor experiência ao cliente.

5.2 Análise de Produtos, Seções e Departamentos Populares

Para visualizar os produtos, seções e departamento populares foram gerados gráficos de contagem horizontal com valores ordenados de forma decrescente. Além disso, foram criados gráficos específicos para detalhar a hierarquia de categorias, permitindo visualizar, por exemplo, quais seções são mais populares dentro de um determinado departamento ou quais produtos são mais comprados dentro de uma seção específica.

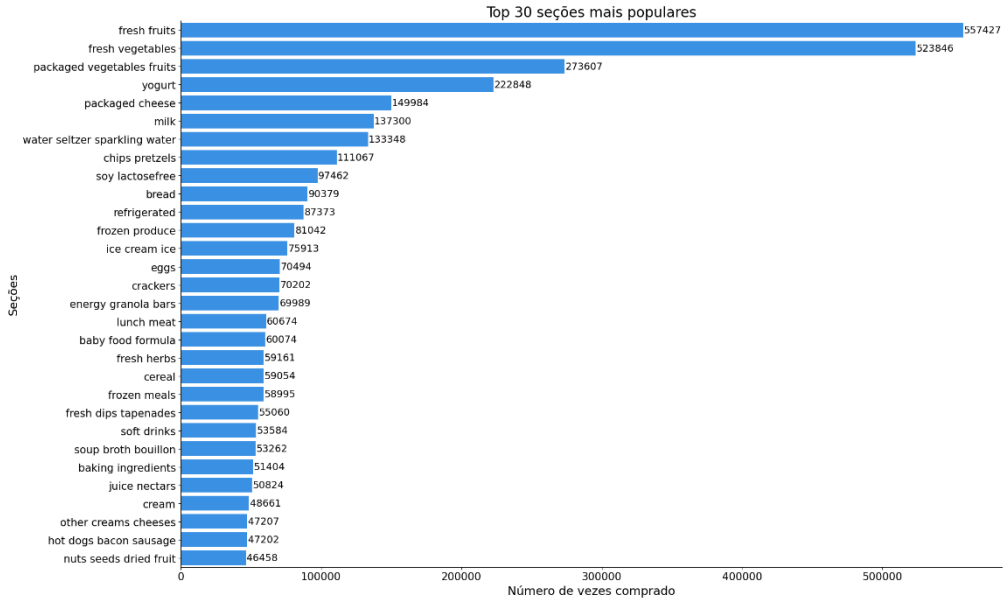

Figura 5. Gráfico de contagem das 30 seções mais populares

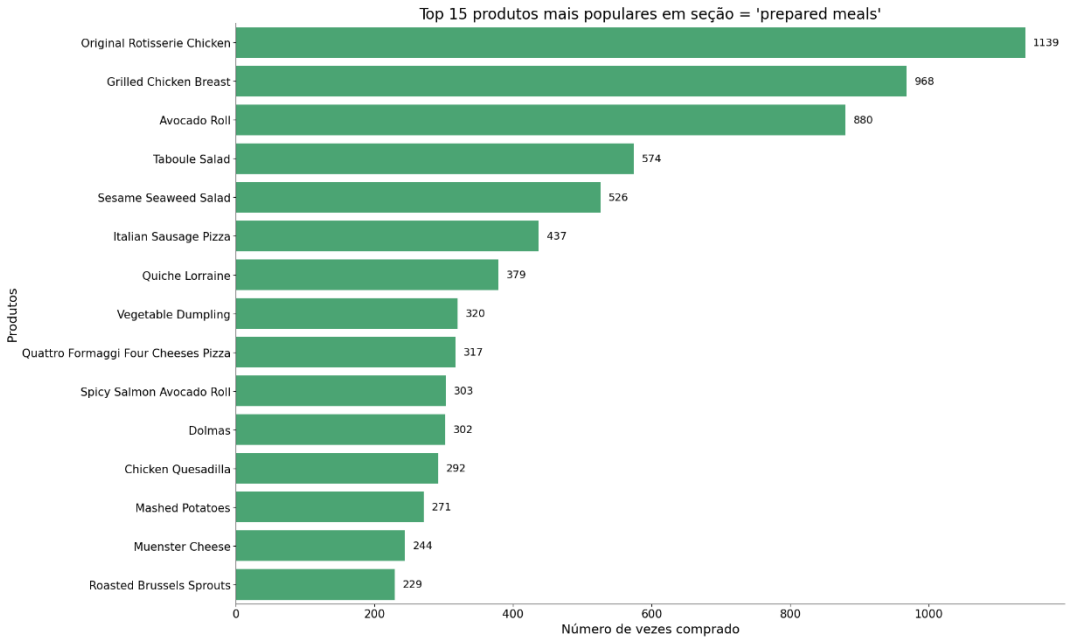

Figura 6. Gráfico de contagem dos 15 produtos mais populares da seção 'prepared meals'

5.3 Análise de Comportamento dos Clientes

Com relação ao comportamento dos clientes foram gerados histogramas da distribuição do número de pedidos por usuário e do intervalo de dias entre pedidos. Para esse último, desconsiderou-se o primeiro pedido de cada usuário (onde 'days_since_prior_order' é nulo).

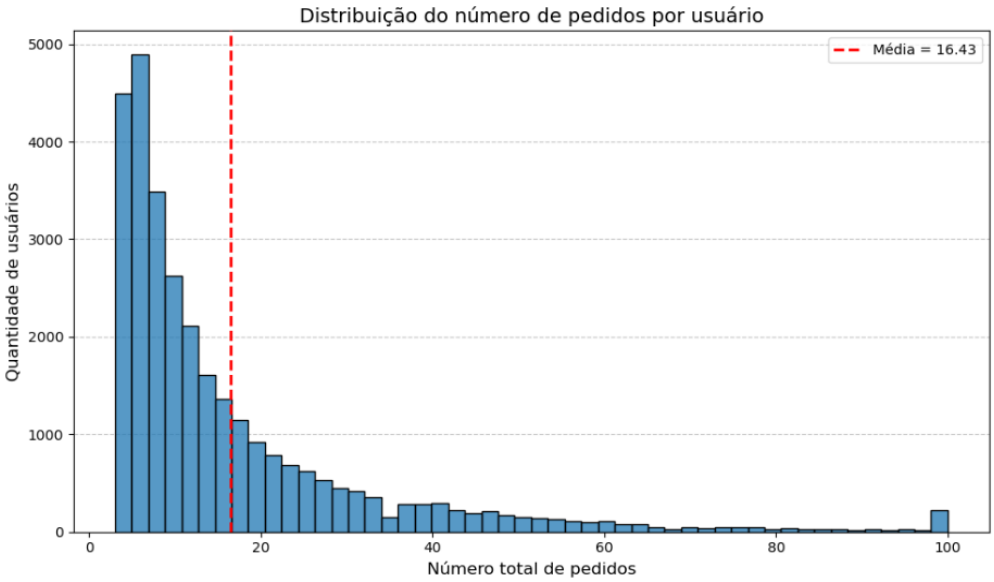

Figura 7. Histograma do número de pedidos por usuário

Percebe-se uma quebra da distribuição natural dos pedidos por usuário no valor 100, logo este valor para 'order_number' parece estar funcionando como um *place holder*, indicando que o sistema limitou ou truncou os valores reais maiores que 100. Assim, o valor da média do número total de pedidos é maior que 16,43.

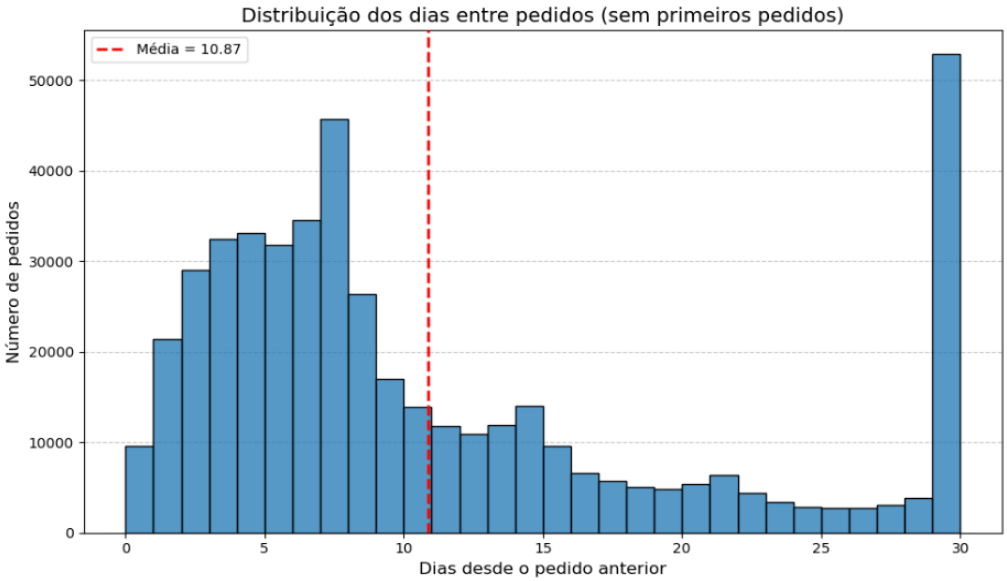

Figura 8. Histograma do intervalo de dias entre pedidos dos usuários

Percebe-se também uma quebra da distribuição natural do intervalo de dias entre pedidos dos usuários no valor 30, logo este valor 'days_since_prior_order' também parece estar

funcionando como um *place holder*. Assim, o valor da média de dias desde o pedido anterior é maior que 10,87.

Outro ponto de interesse foi a análise do engajamento dos clientes. Para isso, avaliou-se se o intervalo médio de dias entre pedidos subsequentes diminui à medida que os usuários realizam mais compras.

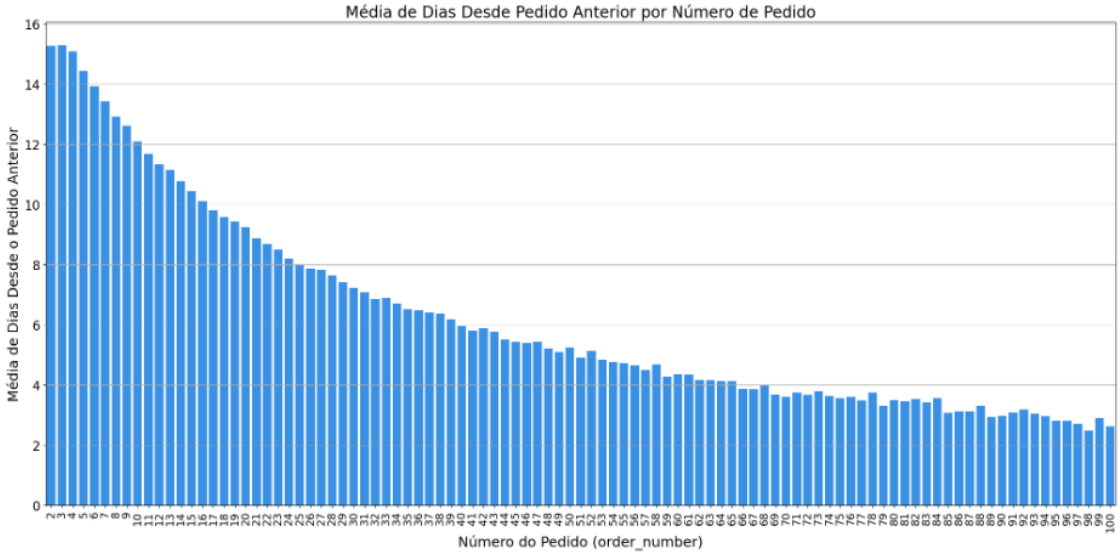

Figura 9. Gráfico de contagem do intervalo médio de dias entre pedidos dos usuários

Dado o gráfico acima, observa-se que, quanto mais o usuário compra, maior é seu engajamento, retornando às compras em intervalos de tempo cada vez menores.

Outra forma de estudar o comportamento dos clientes é ranqueando os departamentos com base na quantidade de pedidos em que o primeiro produto adicionado ao carrinho pertence a eles ('add_to_cart_order' = 1).

produce	138004
dairy eggs	99847
beverages	62820
snacks	34394
frozen	27426
pantry	23903
bakery	18057
deli	13568
household	12810
meat seafood	11191
canned goods	9725
breakfast	8781
personal care	8161
dry goods pasta	7926
alcohol	4990
babies	4044
international	3124
pets	1851
missing	901
other	758
bulk	536

Figura 10. Ranking dos departamentos baseado na quantidade de pedidos em que o primeiro produto adicionado ao carrinho ('add_to_cart_order' igual a 1) pertence a eles

Também foram ranqueados os departamentos e seções com base na média da posição em que seus produtos são adicionados ao carrinho ('add_to_cart_order') nas compras em que

estão presentes. Quanto menor a média, seus produtos frequentemente são os primeiros a serem adicionados no carrinho em compras que os contém.

department	
alcohol	5.431229
beverages	6.978628
dairy eggs	7.460529
pets	7.646084
produce	7.993404
bakery	8.043255
other	8.324291
bulk	8.419398
household	8.453384
meat seafood	8.463264
personal care	8.608231
deli	8.624801
frozen	8.972513
breakfast	9.068645
snacks	9.127630
missing	9.312473
pantry	9.575020
international	9.769304
canned goods	9.850057
dry goods pasta	10.172353
babies	10.295372

aisle	
spirits	4.647271
packaged produce	5.125492
specialty wines champagnes	5.228241
beers coolers	5.258535
milk	5.532484
white wines	5.639490
water seltzer sparkling water	6.145417
red wines	6.146463
soft drinks	6.419248
eggs	6.423923
soy lactosefree	6.732665
energy sports drinks	6.813083
cream	6.866731
fresh fruits	7.128028
cold flu allergy	7.275249
digestion	7.275304
coffee	7.276341
refrigerated	7.278084
cat food care	7.386892
muscles joints pain relief	7.394619

Figura 11. Ranking dos departamentos (todos) e seções (top 20) baseados na média da posição em que seus produtos são adicionados ao carrinho ('add_to_cart_order') nas compras em que estão presentes

5.4 Análise de Recompras

Foi analisado também as seções e departamentos com maior recompra pelos usuários a partir da média de ‘reordered’, ou seja, quanto mais próxima de 1, mais recomprado é a seção/departamento. A imagem a seguir mostra o ranking das seções e departamentos mais comprados pelos usuários, mostrando fidelidade do cliente quanto à eles.

department	
dairy eggs	0.672849
beverages	0.655434
produce	0.651512
bakery	0.629649
deli	0.611725
pets	0.595870
babies	0.587559
snacks	0.575335
meat seafood	0.571917
alcohol	0.569588
bulk	0.565947
breakfast	0.564572
frozen	0.544469
dry goods pasta	0.465012
canned goods	0.457971
missing	0.411659
other	0.407420
household	0.402710
international	0.371340
pantry	0.349842
personal care	0.322329

aisle	
milk	0.784392
water seltzer sparkling water	0.735444
fresh fruits	0.720636
eggs	0.709266
soy lactosefree	0.694301
yogurt	0.691978
packaged produce	0.690757
cream	0.687512
bread	0.674792
refrigerated	0.662836
white wines	0.658521
breakfast bakery	0.653317
energy sports drinks	0.649704
packaged vegetables fruits	0.642279
frozen breakfast	0.633728
soft drinks	0.633006
bulk dried fruits vegetables	0.632696
cat food care	0.629622
prepared meals	0.627233
baby accessories	0.621693

Figura 13. Ranking dos departamentos (todos) e seções (top 20) mais recomprados

Para o ranqueamento dos produtos, optou-se por considerar apenas aqueles cuja quantidade de pedidos é maior ou igual à média geral de pedidos por produto, a fim de evitar que itens com poucas vendas, mas alta taxa de recompra, distorçam o desempenho no ranking.

product_name	
Whole Wheat Multigrain Pop Cakes	0.900709
Organic Lactose Free Whole Milk	0.884265
Porcini & Truffle Ravioli In Egg Pasta	0.883436
Organic Homogenized Whole Milk	0.882581
Organic Large Grade AA Eggs	0.873016
100% Spring Water	0.872832
Original Acai Juice	0.871212
Water Mineral	0.870283
Regular Cream Cheese Spread	0.867550
Organic Whole Milk	0.863902
Food for Cats, Chicken & Herring Formula	0.860927
Farmer Cheese No Salt Added	0.860870
Yerba Mate Sparkling Classic Gold	0.860169
Rose Black	0.858757
Organic Reduced Fat Omega-3 Milk	0.858650
Healthy Kids Organic Vanilla Nutritional Shake	0.858156
Cranberry Pomegranate Sparkling Yerba Mate	0.854701
Organic Reduced Fat Milk	0.854373
Milk, Organic, Vitamin D	0.854252
Whole Organic Omega 3 Milk	0.852836

Figura 14. Ranking dos produtos (top 20) mais recomprados que possuem quantidade de pedidos maior ou igual a média da quantidade de pedidos por produto em geral

Além disso, foi analisada a proporção de produtos com nenhuma recompra entre todos os produtos e foi aferido que 26,94% dos produtos nunca foram adquiridos novamente. Em um cenário de compras presenciais, esse resultado pode indicar um impacto negativo na gestão de estoque e na ocupação de espaço em loja, comprometendo a eficiência operacional e a alocação de recursos, além de gerar imobilização de capital em estoque parado e aumento custo de armazenagem, exposição e controle.

Por fim, foi gerado um gráfico de dispersão entre a taxa de recompra e o número de vendas por seção, com o objetivo de aprofundar a análise da eficiência e atratividade da diversidade de produtos.

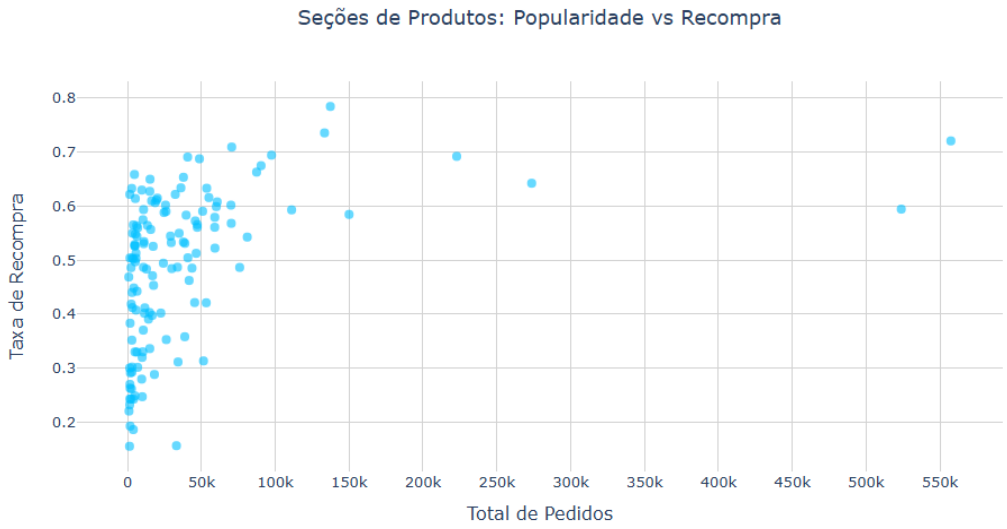

Figura 15. Gráfico de dispersão entre a taxa de recompra e o número de vendas por seções de produtos

Desse gráfico conclui-se que, um número pequeno de categorias de produto responde por alta fidelidade dos clientes. Além disso, observa-se uma concentração na região superior esquerda do gráfico: categorias que apresentam alta taxa de recompra, mas que não possuem um valor expressivo no total de pedidos, indicando alta satisfação, mesmo sem popularidade massiva, esses produtos são bons candidatos a se tornarem essenciais para

nichos específicos. Por fim, nota-se a presença de muitas categorias com baixa taxa de recompra e baixo volume de pedidos, recomendando-se avaliar sua retirada ou investigar as razões de sua baixa performance.

6. Segmentação de Clientes

A segmentação de clientes foi realizada com base em uma matriz de proporções, que representa a fração do total de compras de cada usuário atribuída a cada departamento de produtos. Essa abordagem permitiu agrupar consumidores com padrões de compra semelhantes.

department	alcohol	babies	bakery	beverages	breakfast	bulk	canned goods	dairy eggs	deli	dry goods pasta	...	household	international	meat seafood	missing
user_id															
1	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.021739	0.195652	0.021739	0.021739	...	0.021739	0.086957	0.000000	0.00000
2	0.0	0.0	0.023256	0.260465	0.000000	0.0	0.027907	0.153488	0.060465	0.000000	...	0.004651	0.000000	0.041860	0.00000
3	0.0	0.0	0.174419	0.023256	0.023256	0.0	0.058140	0.348837	0.000000	0.034884	...	0.000000	0.011628	0.000000	0.05814
4	0.0	0.0	0.057692	0.057692	0.019231	0.0	0.153846	0.288462	0.000000	0.000000	...	0.134615	0.000000	0.019231	0.00000
5	0.0	0.0	0.000000	0.108108	0.000000	0.0	0.027027	0.243243	0.000000	0.135135	...	0.054054	0.000000	0.000000	0.00000

Figura 16. Recorte da matriz de proporções dos usuários

6.1 Escolha do Algoritmo de Clusterização

Como observado na análise exploratória (EDA), o conjunto de dados apresenta um forte desbalanceamento nas ocorrências dos valores da variável 'department'. Diante disso, foi feito um estudo comparativo entre o desempenho dos algoritmos de clusterização K-means (que assume clusters esféricos) e Gaussian Mixture Model (que permite clusters com diferentes formatos).

Para escolha do modelo mais adequado decidiu-se avaliar as métricas Silhouette Score e Davies-Bouldin Index em um intervalo de 2 a 10 clusters, a fim de explorar diferentes granularidades de agrupamento. Essa análise visa avaliar os clusters formados pelos diferentes modelos quanto à sua coesão (quão compactos são) e separação (quão distantes estão uns dos outros), permitindo selecionar a abordagem que melhor segmenta os clientes com base em seus padrões de compra.

- **Silhouette Score**: Valores dentro do intervalo [1, -1]. Mede a coesão (distância dentro do cluster) e a separação (distância para outros clusters) por ponto. Valores próximos de 1 indicam bons clusters, enquanto valores negativos sugerem erros de agrupamento.
- **Davies-Bouldin Index**: Valores dentro do intervalo [0, ∞). Mede a qualidade dos clusters com base na razão entre a dispersão dentro dos clusters e a separação entre eles. Valores menores indicam melhor qualidade.

O K-means apresentou Silhouette Scores superiores para todos os números de clusters testados, indicando maior coesão e melhor separação entre os grupos. Além disso, obteve índices de Davies-Bouldin mais baixos, reforçando a qualidade dos agrupamentos em termos de dispersão interna e distinção entre clusters. Por outro lado, o GMM teve desempenho insatisfatório, com Silhouette Scores negativos e índices de Davies-Bouldin elevados, sugerindo que os dados não se ajustam bem à suposição de distribuições gaussianas. Assim, o K-means se mostrou o algoritmo mais apropriado para a segmentação dos clientes neste conjunto de dados.

Tabela 6. Silhouette Scores obtidos

Número de Clusters	GMM	K-means
2	0,016	0,227
3	-0,054	0,200
4	-0,068	0,154
5	-0,066	0,161
6	-0,075	0,159
7	-0,100	0,123
8	-0,109	0,127
9	-0,083	0,129
10	-0,087	0,121

Tabela 7. Índices de Davies-Bouldin obtidos

Número de Clusters	GMM	K-means
2	5,436	1,504
3	7,980	1,774
4	7,292	1,746
5	6,632	1,665
6	7,679	1,656
7	7,329	1,794
8	6,782	1,790
9	6,912	1,731
10	6,569	1,731

6.2 Escolha do Número de Clusters

A escolha do número de clusters foi orientada pela busca por uma segmentação eficaz para impulsionar vendas cruzadas e aprimorar a experiência digital do usuário, equilibrando a interpretabilidade dos perfis gerados com a eficiência computacional do modelo.

Para identificar o ponto de melhor equilíbrio entre a compactação dos clusters e a complexidade do modelo, foi utilizado o Elbow Method (Método do Cotovelo) no intervalo de 2 a 20 clusters. Essa abordagem se baseia na análise da inércia, que representa a soma das distâncias quadradas entre cada ponto e o centróide do seu respectivo cluster. Valores menores de inércia indicam maior coesão interna, ou seja, pontos mais próximos entre si dentro de um mesmo grupo.

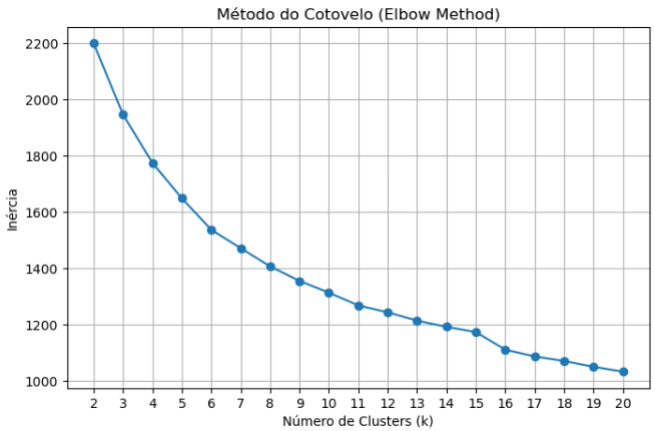

Figura 16. Curva do Elbow Method

Tabela 8. Diferenças relativas e absolutas de inércia entre números de clusters subsequentes

Transição de Número de Clusters	Diferença Absoluta	Diferença Relativa (%)
2 → 3	-253,47	-11,52
3 → 4	-171,78	-8,83
4 → 5	-125,60	-7,08
5 → 6	-111,87	-6,78
6 → 7	-65,39	-4,25
7 → 8	-64,62	-4,39
8 → 9	-51,86	-3,69
9 → 10	-41,55	-3,07
10 → 11	-45,37	-3,45
11 → 12	-24,53	-1,93
12 → 13	-29,54	-2,37
13 → 14	-22,11	-1,82
14 → 15	-18,65	-1,56
15 → 16	-62,48	-5,32
16 → 17	-24,04	-2,16
17 → 18	-16,21	-1,49
18 → 19	-20,55	-1,92
19 → 20	-17,16	-1,63

Analisando o gráfico e a tabela, percebe-se que, a partir de 4 clusters a inércia deixa de decrescer com a mesma intensidade. Esse comportamento se acentua ainda mais após 6 clusters, indicando que o acréscimo de novos grupos além desse ponto representa alto custo computacional com pouca melhoria na qualidade do clustering.

Para seguir com a escolha do número de clusters foram analisadas novamente as métricas Silhouette Score e Davies-Bouldin Index.

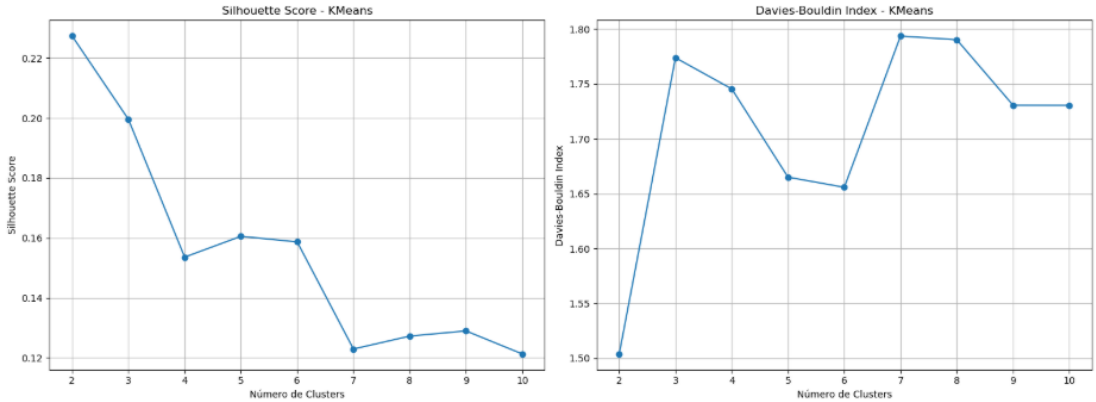

Figura 17. Gráficos de Silhouette Score e Davies-Bouldin Index para K-means

A partir dos gráficos apresentados, observa-se que, a utilização de mais de 6 clusters resulta em menor separação entre os grupos, evidenciada pela queda acentuada no Silhouette Score e pelo aumento no Davies-Bouldin Index. Esse comportamento indica que a segmentação perde eficácia com um número excessivo de grupos, dificultando a definição de perfis distintos e, consequentemente, reduzindo o potencial de ações de vendas personalizadas.

Considerando o contexto de segmentação de clientes de um supermercado, onde se busca ações de marketing mais diretas e operacionais, optou-se por um número de clusters igual

a 3. Com o modelo definido, foi possível classificar os 30.000 clientes em 3 grupos distintos (clusters), numerados de 0 a 2.

6.3 Análise de Perfil dos Grupos de Clientes

O Silhouette score de aproximadamente 0,20 e o índice de Davies-Bouldin de aproximadamente 1,77 indicam clusters com alguma sobreposição, o que é esperado em dados de hábitos de compra devido à sua natureza heterogênea. Para validar o uso de 3 clusters foi avaliada a interpretabilidade dos perfis, e, para isso, foram analisados os centróides dos 3 clusters, que representam a média das proporções de compra por departamento entre os usuários de cada grupo. Esses centróides indicam padrões de consumo típicos e permitem identificar preferências e tendências de compra distintas, definindo o perfil de consumo de cada grupo.

A figura abaixo apresenta uma tabela resumo dos três centróides.

	0	1	2
alcohol	0.007226	0.001836	0.027901
babies	0.013055	0.005301	0.005489
bakery	0.043129	0.026163	0.026923
beverages	0.073024	0.051947	0.280881
breakfast	0.026633	0.013834	0.024962
bulk	0.000749	0.001230	0.000793
canned goods	0.037072	0.032551	0.016853
dairy eggs	0.193316	0.132189	0.093054
deli	0.035160	0.028180	0.022687
dry goods pasta	0.031441	0.021964	0.013593
frozen	0.091607	0.046779	0.058320
household	0.029467	0.012208	0.066033
international	0.009432	0.007553	0.005180
meat seafood	0.025466	0.023010	0.010623
missing	0.002681	0.002334	0.002614
other	0.001470	0.000837	0.002446
pantry	0.070566	0.052909	0.047186
personal care	0.018368	0.008494	0.032833
pets	0.004375	0.000885	0.005485
produce	0.198969	0.474730	0.080304
snacks	0.086790	0.055066	0.175841

Figura 18. Centróides dos clusters

A fim de facilitar a diferenciação e classificação dos perfis, foi construída uma tabela que mostra, para cada categoria de um centroide, a diferença relativa entre seu valor e a média dos valores dessa categoria nos outros dois centroides.

	0	1	2
alcohol	-0.513981	-0.895446	5.157249
babies	1.419702	-0.428266	-0.401907
bakery	0.624875	-0.253044	-0.222910
beverages	-0.561188	-0.706437	3.495133
breakfast	0.372968	-0.463755	0.233717
bulk	-0.259223	0.595157	-0.198785
canned goods	0.500786	0.207263	-0.515883
dairy eggs	0.716521	-0.076799	-0.428251
deli	0.382408	-0.025694	-0.283641
dry goods pasta	0.768484	-0.024566	-0.490941
frozen	0.743242	-0.375978	-0.157133
household	-0.246749	-0.744344	2.168945
international	0.481467	0.033839	-0.390052
meat seafood	0.514359	0.275203	-0.561740
missing	0.083861	-0.118489	0.042371
other	-0.104409	-0.572610	1.120613
pantry	0.409982	-0.101349	-0.235700
personal care	-0.111088	-0.668190	1.444495
pets	0.373754	-0.820464	1.085285
produce	-0.283036	2.399748	-0.761603
snacks	-0.248262	-0.580662	1.479141

Figura 19. Diferença relativa entre valores dos centróides

Pelas tabelas é possível identificar as principais características de cada grupo de clientes.

- **Cluster 0:**
  - Alta proporção de ‘produce’ (0,198969) e ‘dairy eggs’ (0,193316), indicando preferência por laticínios e produtos frescos.
  - Proporções moderadas em ‘frozen’ (0,091607), ‘snacks’ (0,08679), ‘beverages’ (0,073024) e ‘pantry’ (0,070566).
  - Baixa proporção em ‘bulk’ (0,000749), ‘other’ (0,001470) e ‘missing’ (0,002681), mostrando pouco interesse em compras em grandes quantidades ou itens diversos.
  - Elevada proporção de babies (0,013055) em comparação com o cluster 1 e 2 (0,005301 e 0,005489 respectivamente).
- **Cluster 1:**
  - Altíssima proporção em ‘produce’ (0,47473) e alta proporção em ‘dairy eggs’ (0,132189), indicando forte preferência por produtos frescos e laticínios.
  - Proporções moderadas em ‘snacks’ (0,05506), ‘pantry’ (0,052909), ‘beverages’ (0,051947) e ‘frozen’ (0,046779).
  - Baixa proporção em ‘alcohol’ (0,001836), pets (0,000885) e ‘other’ (0,000837), indicando pouco interesse em bebidas alcoólicas, itens para pets e produtos diversos.
- **Cluster 2:**
  - Alta proporção em ‘beverages’ (0,280881) e ‘snacks’ (0,175841), mostrando uma preferência dominante por bebidas e ‘snacks’.
  - Proporções moderadas em ‘diary eggs’ (0,093054), ‘produce’ (0,080304) e ‘household’ (0,066033).
  - Baixa proporção em ‘bulk’ (0.000793), ‘babies’ (0,005489) e ‘pets’ (0,005485), indicando pouco interesse em compras em grandes quantidades, produtos internacionais e itens para bebês.
  - Elevada proporção de ‘alcohol’ (0,027901) em comparação com o cluster 0 e 1 (0,007226 e 0,001836 respectivamente)

- Elevada proporção de ‘personal care’ (0,032833) em comparação com o cluster 0 e 1 (0,018368 e 0,008494 respectivamente)

De acordo com essas características podemos resumir os perfis dos grupos de clientes em:

- **Cluster 0:** Consumidores variados, focados em laticínios e produtos frescos, com compras planejadas (possivelmente famílias).
- **Cluster 1:** Consumidores variados, focados em saúde, com ênfase em produtos frescos e laticínios (indivíduos preocupados com nutrição, alguns possivelmente vegetarianos, atletas e fitness).
- **Cluster 2:** Consumidores que buscam praticidade, priorizando snacks e bebidas (jovens, solteiros ou casais sem filhos com estilo de vida casual e ativa).

Os centróides dos clusters revelam perfis claramente distintos, oferecendo insights acionáveis para estratégias de marketing descritas adiante. Logo, a adoção de 3 clusters equilibra qualidade de segmentação, interpretabilidade dos perfis e utilidade prática.

O gráfico a seguir mostra a proporção de usuários em cada perfil de cliente no conjunto de dados:

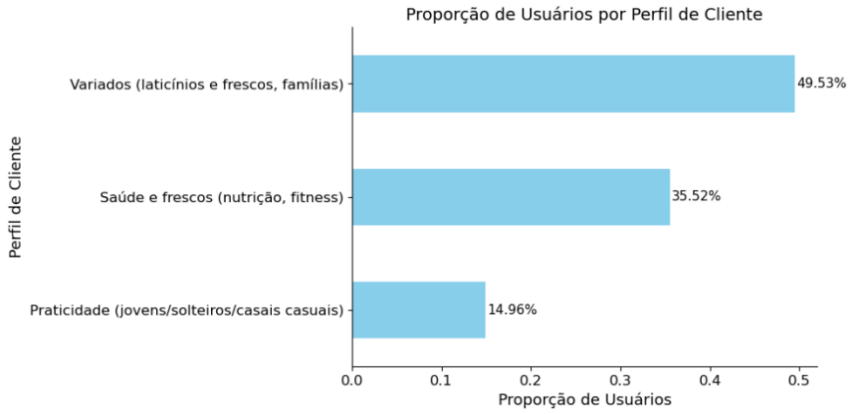

Figura 20. Proporção de usuários por perfil de cliente

6.4 Visualização 2D por PCA

Para viabilizar a visualização da separação entre os grupos de clientes em um espaço bidimensional, aplicou-se a Análise de Componentes Principais (PCA). Para isso, os dados foram padronizados utilizando o StandardScaler, garantindo média zero e variância unitária para todas as variáveis, assegurando que a projeção do PCA não fosse enviesada por atributos com maior variabilidade.

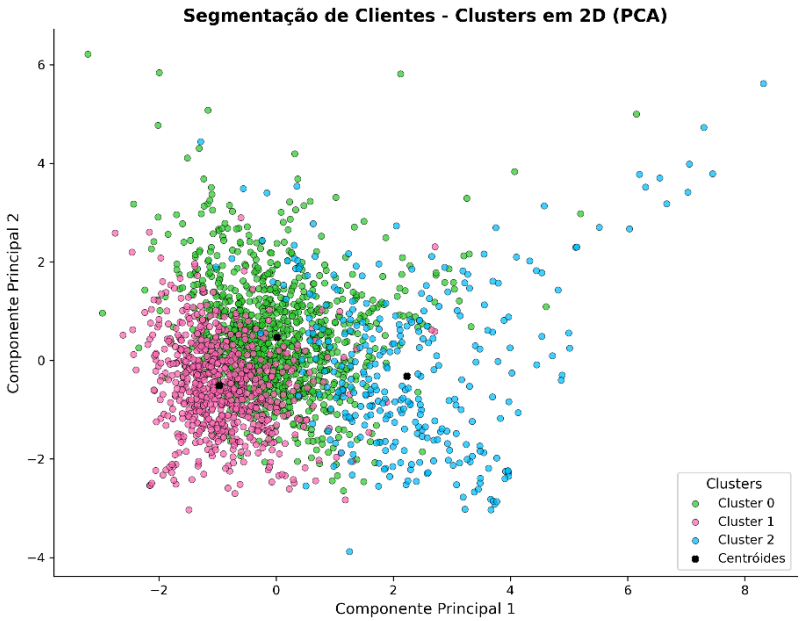

Figura 21. Visualização 2D dos clusters por PCA

A visualização acima, embora uma simplificação 2D dos dados de alta dimensionalidade, corrobora com os valores de Silhouette Score de Davies-Bouldin Index obtidos, indicando separação entre clusters, mas com certa sobreposição entre eles.

Além disso, percebe-se uma maior dispersão do Cluster 0, o que é coerente com seu perfil de consumidores variados e famílias no geral, que não apresentam um padrão específico de compra, refletindo escolhas mais heterogêneas.

Por outro lado, a maior coesão visual dos Cluster 1 indica um grupo de consumidores com hábitos de compra mais consistentes, fortemente alinhados ao perfil com foco em saúde. Esse padrão é reforçado pela alta concentração de compras no departamento 'produce' (frutas, legumes e verduras), que corresponde a cerca de 50% do volume de compras desses consumidores.

Já o Cluster 2 apresenta uma maior distinção espacial em relação aos demais, especialmente nos extremos da projeção, sugerindo um perfil mais específico. Essa diferenciação é justificada pela predominância combinada de compras nos departamentos de 'beverages' e 'snacks', em contraste com os demais grupos, que apresentam proporções significativamente menores de compras nesses departamentos.

7. Regras de Associações

Com o objetivo de impulsionar vendas cruzadas por meio de recomendações personalizadas, foram extraídas regras de associação entre as categorias (seções) dos produtos para cada cluster de clientes, considerando os departamentos mais comprados. O objetivo foi identificar quais combinações de categorias de produtos apresentam associação positiva, ou seja, tendem a ser compradas em conjunto, e quais apresentam associação negativa, indicando menor probabilidade de ocorrência conjunta.

As regras foram extraídas utilizando o algoritmo Apriori, configurado com um suporte mínimo de 0,02 Esse valor foi definido considerando o tamanho do conjunto de dados e a alta diversidade de seções de produtos (134 ao todo), o que naturalmente reduz a

frequência individual de muitas combinações. Assim, o suporte de 2% permite capturar padrões relevantes e recorrentes, ao mesmo tempo em que evita incluir associações raras ou pouco representativas.Os departamentos selecionados para a extração das regras de associação de cada cluster foram definidos com base no perfil de consumo dos grupos de clientes, considerando aqueles com maior proporção de compras em relação aos demais departamentos.

- **Cluster 0:** 'dairy eggs', 'produce', 'frozen', ‘snacks’.
- **Cluster 1:** 'produce', 'dairy eggs'.
- **Cluster 2:** 'beverages', 'snacks'.

A seguir, os gráficos de dispersão permitem visualizar a relação entre o suporte e o lift das regras de associação extraídas para cada cluster de clientes, evidenciando os padrões de comportamento de compra e a força das associações entre categorias de produtos em cada grupo.

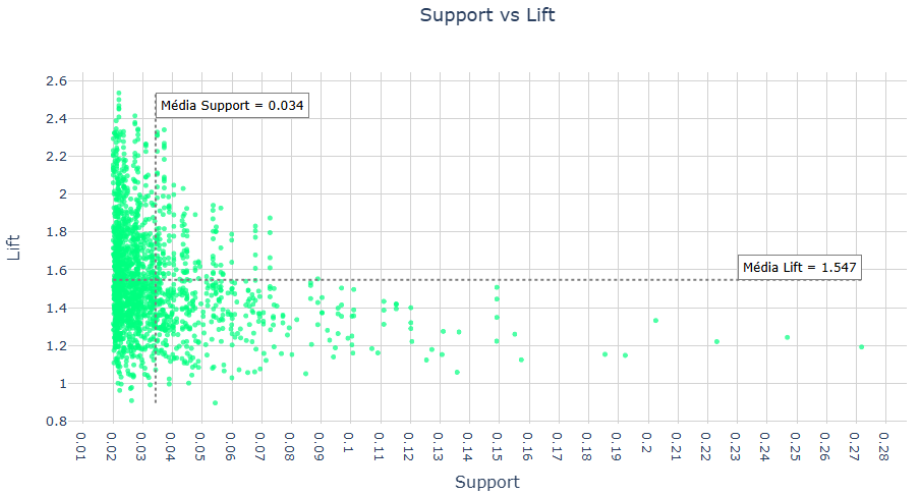

Figura 22. Gráfico de dispersão entre support e lift das regras de associação do cluster 0

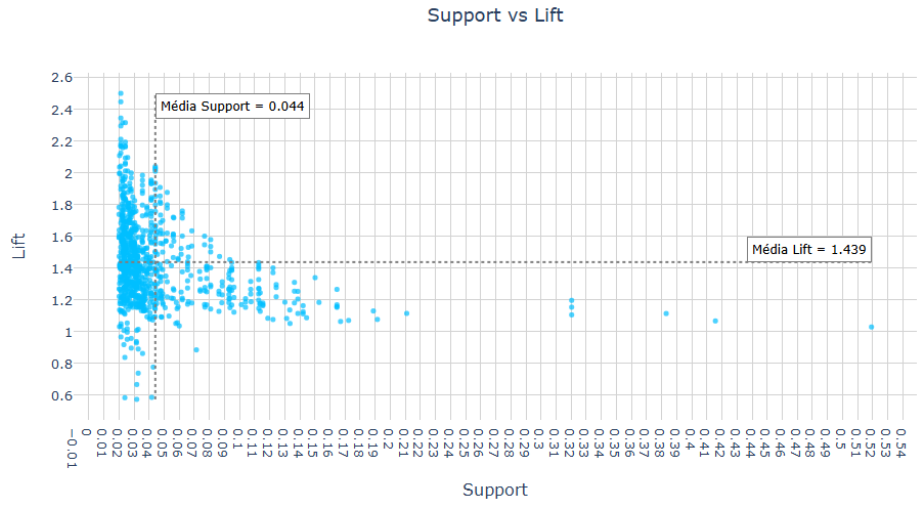

Figura 23. Gráfico de dispersão entre support e lift das regras de associação do cluster 1

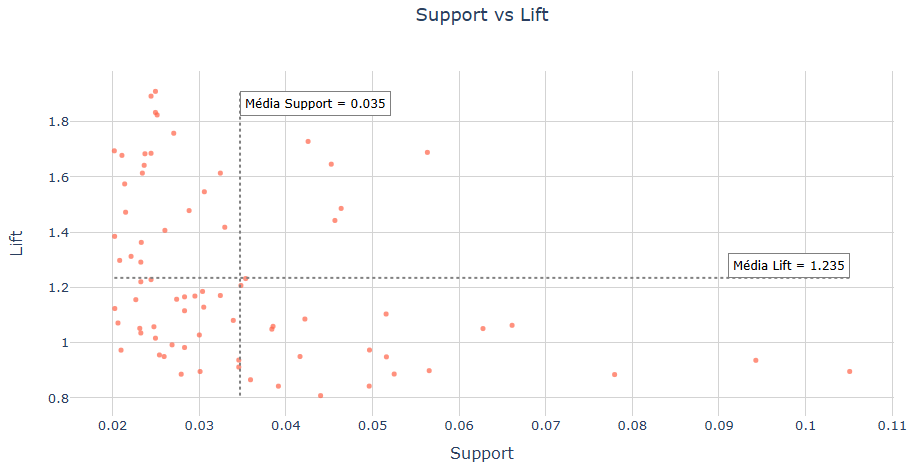

Figura 24. Gráfico de dispersão entre support e lift das regras de associação do cluster 2

Observa-se que, de modo geral, quanto maior o suporte de uma regra, menor tende a ser o seu lift, indicando que as categorias mais frequentes de produtos tendem a ser compradas em conjunto naturalmente. Por outro lado, regras com suporte baixo e lift elevado indicam associações menos comuns, e mais significativas, apontando que as categorias associadas são compradas em conjunto de forma não aleatória. Uma análise mais aprofundada das regras específicas de cada cluster permite identificar associações exclusivas entre categorias, e, com base nessas informações, é possível propor sistemas de recomendação mais eficientes e com melhor retorno, estimulando o consumo cruzado de categorias com forte associação positiva e evitando combinações ineficazes.

Com isso, a fim de maximizar o impacto das estratégias de marketing e influenciar positivamente as decisões de compra dos usuários em ambientes digitais, foram selecionadas, para cada cluster, associações fortes, com alto lift e nível de confiança acima da média e alinhadas ao seu perfil de consumo. O uso de regras com alta confiança visa garantir que as recomendações representem padrões de compra consistentes, aumentando a relevância das sugestões e a probabilidade de conversão. Paralelamente, foram desconsideradas regras com suporte alto e lift baixo, que refletem padrões naturais de compra sem potencial incremental, e regras com lift abaixo de 1, que indicam menor probabilidade de ocorrência conjunta. As regras selecionadas para cada grupo de usuários estão apresentadas a seguir.

Tabela 9. Regras de Associação Selecionadas - Cluster 0 (Consumidores Variados)

Antecedents	Consequents	Support	Confidence	Lift
Milk, Fresh Fruits, Fresh Vegetables	Yogurt, Packaged Cheese	0,028	0,283	2,346
Fresh Fruits, Packaged Cheese	Yogurt, Fresh Vegetables	0,056	0,303	1,927
Chips/Pretzels, Fresh Vegetables	Fresh Fruits, Packaged Vegetables/Fruits	0,045	0,446	1,806
Milk, Frozen Produce	Yogurt	0,025	0,523	1,560

Tabela 10. Regras de Associação Fortes - Cluster 1 (Nutrição e Fitness)

Antecedents	Consequents	Support	Confidence	Lift
Packaged Cheese, Yogurt	Milk, Fresh Vegetables	0,026	0,350	2,098
Milk, Fresh Vegetables	Fresh Fruits, Packaged Cheese	0,049	0,293	1,702
Milk, Fresh Herbs	Fresh Fruits, Fresh Vegetables	0,034	0,795	1,531
Fresh Fruits, Soy Lactosefree	Packaged Vegetables/Fruits, Fresh Vegetables	0,075	0,541	1,414

Tabela 11. Regras de Associação Fortes - Cluster 2 (Praticidade)

Antecedents	Consequents	Support	Confidence	Lift
Nuts/Seeds/Dried/Fruit	Energy/Granola/Bars	0,025	0,250	1,824
Cookies/Cakes	Chips/Pretzels	0,043	0,399	1,728
Crackers	Chips/Pretzels	0,056	0,390	1,689
Refrigerated	Juice/Nectars	0,045	0,224	1,646

8. Estratégias de Vendas Segmentadas

8.1 Sistemas de Recomendação e Personalização Digital

Com os perfis dos usuários bem definidos e as fortes regras de associação selecionadas para cada cluster é possível criar sistemas de recomendação estratégicos para impulsionar vendas cruzadas com base no carrinho digital do usuário e seu respectivo grupo.

Tabela 12. Sistema de Recomendação - Cluster 0 (Consumidores Variados)

Gatilho (Antecedente no carrinho)	Recomendação (Consequente sugerido)
Milk + Fresh Fruits + Fresh Vegetables	Yogurt e Packaged Cheese
Fresh Fruits + Packaged Cheese	Yogurt e Fresh Vegetables
Chips/Pretzels + Fresh Vegetables	Fresh Fruits e Packaged Vegetables/Fruits
Milk + Frozen Produce	Yogurt

Tabela 13. Sistema de Recomendação - Cluster 1 (Nutrição e Fitness)

Gatilho (Antecedente no carrinho)	Recomendação (Consequente sugerido)
Packaged Cheese + Yogurt	Milk e Fresh Vegetables
Milk + Fresh Vegetables	Fresh Fruits e Packaged Cheese
Milk + Fresh Herbs	Fresh Fruits e Fresh Vegetables
Fresh Fruits + Soy LactoseFree	Packaged Vegetables/Fruits e Fresh Vegetables

Tabela 14. Sistema de Recomendação - Cluster 2 (Praticidade)

Gatilho (Antecedente no carrinho)	Recomendação (Consequente sugerido)
Nuts/Seeds/Dried Fruit	Energy/Granola/Bars
Cookies/Cakes	Chips/Pretzels
Crackers	Chips/Pretzels
Refrigerated	Juice/Nectars

A ideia é que, ao identificar os antecedentes no carrinho do usuário, o sistema exiba sugestões de compra com base nos consequentes (seções de produtos). Caso o usuário siga a recomendação, será direcionado para a seção correspondente.

Combinada ao sistema de recomendação, uma seção “Para Você”, contendo uma ordenação estratégica de produtos também contribui para impulsionar as vendas. Essa ordenação considera as seções mais populares dentro dos departamentos mais representativos de cada grupo de clientes. Para isso, foram selecionadas as três seções mais populares de cada um dos departamentos mais representativos e adotou-se um score que combina a proporção de compras de cada produto em sua respectiva seção (‘aisle’) com sua taxa média de recompra, seguindo a fórmula: score = proporcao × (taxa de recompra) ^{2} × 1000.

A elevação ao quadrado da taxa de recompra foi utilizada para ampliar o peso da fidelidade do consumidor, priorizando produtos que, além de populares em volume, apresentam alto índice de recorrência. Com isso, valorizam-se produtos que proporcionam uma experiência positiva, e geram retornos consistentes, sendo ao mesmo tempo relevantes em frequência e retenção.

Com os scores calculados para os produtos, são selecionados os 10 produtos com maior pontuação em cada seção, gerando uma lista personalizada “Para Você” por categoria de produtos. Assim, ao entrar em uma dessas seções de produtos, o cliente se depara inicialmente aos produtos mais relevantes para o seu perfil de consumo.

Os departamentos mais representativos para cada cluster, conforme visto anteriormente são 'dairy eggs', 'produce', 'frozen' e ‘snacks’ para Cluster 0, ‘produce’ e ‘diary eggs’ para Cluster 1 e 'beverages', 'snacks' para Cluster 2.

As três seções mais populares, em ordem, desses departamentos são apresentadas na tabela a seguir:

Tabela 15. Três seções de produtos mais populares por departamento - Cluster 0

Departamento	Seções Populares
Dairy Eggs	Yogurt, Packaged Cheese, Milk
Produce	Fresh Fruits, Fresh Vegetables, Packaged Vegetables/Fruits
Frozen	Ice Cream, Frozen Produce, Frozen Meals
Snacks	Chips/Pretzels, Crackers, Energy/Granola/Bars

Tabela 16. Três seções de produtos mais populares por departamento - Cluster 1

Departamento	Seções Populares
Produce	Fresh Vegetables, Fresh Fruits, Packaged Vegetables/Fruits
Dairy Eggs	Yogurt, Packaged Cheese, Milk

Tabela 17. Três seções de produtos mais populares por departamento - Cluster 2

Departamento	Seções Populares
Beverages	water seltzer sparkling water, soft drinks, refrigerated
Snacks	Chips/Pretzels, Energy/Granola/Bars, Candy/Chocolate

A seguir seguem alguns exemplos das listas personalizadas de produtos (“Para Você”), geradas para cada grupo de clientes.

product_name	count	taxa_recompra	proporcao	score
Lightly Salted Baked Snap Pea Crisps	2575	0.662913	0.039590	17.398131
Corn Chips	841	0.780024	0.012930	7.867278
Sea Salt Pita Chips	1343	0.611318	0.020649	7.716549
Aged White Cheddar Baked Rice & Corn Puffs Glu...	1053	0.683761	0.016190	7.569190
Sea Salt & Vinegar Potato Chips	1231	0.614947	0.018927	7.157255
Sea Salt Potato Chips	1160	0.628448	0.017835	7.043846
Pretzel Crisps Original Deli Style Pretzel Cra...	1363	0.570800	0.020956	6.827727
Veggie Chips	811	0.700370	0.012469	6.116298
Thin & Light Tortilla Chips	848	0.672170	0.013038	5.890697
Organic Tortilla Chips	833	0.651861	0.012807	5.442112

Figura 25. Lista personalizada de Chips/Pretzels do cluster 0

product_name	count	taxa_recompra	proporcao	score
Total 2% with Strawberry Lowfat Greek Strained...	2830	0.772085	0.019799	11.802522
Total 0% Nonfat Greek Yogurt	2405	0.732640	0.016826	9.031401
Total 2% Lowfat Greek Strained Yogurt With Blu...	2049	0.783797	0.014335	8.806584
Total 2% Lowfat Greek Strained Yogurt with Peach	2049	0.782333	0.014335	8.773714
Icelandic Style Skyr Blueberry Non-fat Yogurt	2034	0.783186	0.014230	8.728487
Non Fat Raspberry Yogurt	1825	0.803288	0.012768	8.238790
Total 2% Greek Strained Yogurt with Cherry 5.3 oz	1915	0.780679	0.013398	8.165297
Total 0% Greek Yogurt	2174	0.723551	0.015210	7.962625
Total 2% All Natural Greek Strained Yogurt wit...	1993	0.738585	0.013943	7.606182
Total Greek Strained Yogurt	1990	0.711055	0.013922	7.039117

Figura 26. Lista personalizada de Yogurt do cluster 0

product_name	count	taxa_recompra	proporcao	score
Total Greek Strained Yogurt	1917	0.684924	0.030006	14.076285
Total 0% Greek Yogurt	1815	0.689807	0.028409	13.518009
Total 0% Nonfat Greek Yogurt	1808	0.684735	0.028300	13.268553
Total 2% with Strawberry Lowfat Greek Strained...	1333	0.718680	0.020865	10.776595
Whole Milk Plain Yogurt	1293	0.726991	0.020239	10.696406
Organic Plain Whole Milk Yogurt	993	0.784491	0.015543	9.565471
Total 2% All Natural Low Fat 2% Milkfat Greek ...	1161	0.715762	0.018172	9.310018
Total 2% Lowfat Greek Strained Yogurt with Peach	961	0.742976	0.015042	8.303358
Yobaby Organic Plain Yogurt	719	0.826147	0.011254	7.681123
Plain Greek Yogurt	774	0.784238	0.012115	7.451044

Figura 27. Lista personalizada de Yogurt do cluster 1

product_name	count	taxa_recompra	proporcao	score
Banana	31469	0.858909	0.104682	77.226246
Bag of Organic Bananas	30381	0.848524	0.101062	72.764236
Organic Strawberries	22685	0.798104	0.075462	48.066908
Organic Hass Avocado	18834	0.820803	0.062651	42.209299
Organic Avocado	16226	0.786269	0.053976	33.368880
Large Lemon	13836	0.734244	0.046025	24.813000
Limes	13326	0.720696	0.044329	23.024616
Strawberries	10785	0.725545	0.035876	18.885846
Apple Honeycrisp Organic	7312	0.752735	0.024323	13.781884
Organic Lemon	7795	0.718409	0.025930	13.382826

Figura 28. Lista personalizada de Fresh Fruits do cluster 1

product_name	count	taxa_recompra	proporcao	score
Organic Simply Naked Pita Chips	507	0.747535	0.025164	14.061722
Organic Tortilla Chips	452	0.672566	0.022434	10.147914
Lightly Salted Baked Snap Pea Crisps	468	0.636752	0.023228	9.417914
Dark Chocolate Pretzels with Sea Salt	337	0.747774	0.016726	9.352748
Pub Mix	318	0.767296	0.015783	9.292244
Sea Salt & Vinegar Potato Chips	367	0.653951	0.018215	7.789767
Backyard Barbeque Potato Chips	285	0.701754	0.014145	6.965995
Salt & Pepper Krinkle Chips	266	0.706767	0.013202	6.594807
Pretzel Crisps Original Deli Style Pretzel Cra...	324	0.620370	0.016081	6.188924
Sea Salt Pita Chips	265	0.675472	0.013153	6.001064

Figura 29. Lista personalizada de Chips/Pretzels do cluster 2

product_name	count	taxa_recompra	proporcao	score
100% Raw Coconut Water	841	0.833532	0.044727	31.075126
Organic Raw Kombucha Gingerade	788	0.836294	0.041908	29.310111
Trilogy Kombucha Drink	651	0.814132	0.034622	22.947935
Lemonade	597	0.668342	0.031750	14.182223
Enlightened Organic Raw Kombucha	358	0.851955	0.019040	13.819410
Synergy Organic Kombucha Gingerberry	389	0.807198	0.020688	13.479772
Original Orange Juice	612	0.632353	0.032548	13.014976
Pulp Free Orange Juice	394	0.713198	0.020954	10.658333
Original No Pulp 100% Florida Orange Juice	450	0.644444	0.023932	9.939312
Kombucha, Organic Raw, Citrus	253	0.794466	0.013455	8.492674

Figura 30. Lista personalizada de Refrigerated do cluster 2

Observa-se que, para diferentes grupos de clientes, uma mesma seção de produtos apresenta uma ordenação distinta dos itens por relevância, refletindo as preferências específicas de cada perfil. Essa personalização é essencial para aumentar a efetividade das recomendações, direcionando a atenção do cliente para os produtos mais alinhados ao seu comportamento de compra e, consequentemente, potencializando conversões e fidelização, além de aprimorar a experiência e o engajamento dos usuários com o ambiente de compra digital.

8.2 Campanhas e Promoções por Perfil de Cliente

Por fim, uma outra estratégia de vendas é a criação de campanhas e promoções direcionadas às seções de produtos destacadas nas regras de associação selecionadas para cada cluster.

- **Cluster 0**: Consumidores variados (famílias)
    - **Foco**: Laticínios, produtos frescos, congelados e snacks.
    - **Estratégia**: Promoções de kits familiares com laticínios, vegetais frescos e alimentos congelados, destacando praticidade para refeições caseiras, além de snacks para momentos de lazer. As campanhas podem incluir receitas práticas para famílias, com destaque para ovos, iogurtes e vegetais prontos para cozinhar.

- **Cluster 1**: Consumidores focados em saúde (indivíduos preocupados com nutrição)
    - **Foco**: Produtos frescos e laticínios.
    - **Estratégia**: Campanhas voltadas à alimentação saudável, promovendo cestas com vegetais frescos e laticínios ricos em proteínas (ex.: queijos magros, ovos orgânicos). Materiais como e-books de receitas saudáveis e nutritivas podem ajudar no engajamento.
- **Cluster 2**: Consumidores que buscam praticidade (jovens ou pessoas ativas)
    - **Foco**: Bebidas não alcoólicas e snacks.
    - **Estratégia**: Ofertas de combos prontos para consumo, ideais para rotina de trabalho/estudos. Campanhas descontraídas nas redes sociais, com foco em praticidade e estilo de vida jovem.

9. Conclusão e Considerações Finais

As regras de associação geradas para os três perfis de clientes, identificados via K-means, viabilizaram sistemas de recomendação segmentados e altamente eficientes. Além disso, a estratégia de combinar o volume de vendas com taxas de recompra permitiu a criação de listas personalizadas inteligentes para cada grupo. Estas duas abordagens, juntamente com promoções e campanhas direcionadas impulsionam vendas, aumentam a retenção e fortalece o engajamento dos clientes digitais.

Embora o Silhouette Score de 0,20 e o índice Davies-Bouldin de 1,77 apontam para uma sobreposição moderada entre os clusters, os perfis gerados são claros, interpretáveis e ofereceram insights valiosos para a criação das estratégias de marketing eficientes. Para futuras iterações, recomenda-se explorar algoritmos como HDBSCAN ou Spectral Clustering, além de reavaliar as variáveis selecionadas e aplicar técnicas de redução de dimensionalidade não linear para aprimorar a separação entre os grupos.

Referências

DAVIES, D. L.; BOULDIN, D. W. **A cluster separation measure**. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. PAMI-1, n. 2, p. 224–227, 1979.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. Burlington, MA: Elsevier, 2012.

HARRIS, C. R. et al. **Array programming with NumPy**. Nature, 2020. Disponível em: https://numpy.org/

HERATH, S. **Fundamentals of Associate Rule Mining**. Medium, 2024. Disponível em: https://medium.com/image-processing-with-python/fundaments-of-associate-rule-mining-468801ec0a29

HUNTER, J. D. **Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 2007**. Disponível em: https://matplotlib.org/

JOBLIB Developers. **joblib: running Python functions as pipeline jobs**. Disponível em: https://joblib.readthedocs.io/

KAGGLE. **Instacart Market Basket Analysis**. 2024. Disponível em: https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis

KALOYANOVA, E. **PCA and K-Means Clustering in Python**. 365 Data Science, 2024. Disponível em: https://365datascience.com/tutorials/python-tutorials/pca-k-means/.

MACQUEEN, J. B. **Some methods for classification and analysis of multivariate observations**. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967.

OPENPYXL Contributors. openpyxl: **A Python library to read/write Excel 2010 xlsx/xlsm files**. Disponível em: https://openpyxl.readthedocs.io/

PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 2011. Disponível em: https://scikit-learn.org/

PANDAS Development Team. **pandas: powerful Python data analysis toolkit**. Disponível em: https://pandas.pydata.org/

PLOTLY Technologies Inc. **Plotly: The front-end for ML and data science models**.

Disponível em: https://plotly.com/python/

RASCHKA, S. **Apriori – mlxtend**. Disponível em: https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/#api

RASCHKA, S. mlxtend: **Machine Learning Extensions**. Disponível em: https://rasbt.github.io/mlxtend/

REYNOLDS, D. A. **Gaussian Mixture Models**. In: Encyclopedia of Biometrics. Springer, 2009.

ROUSSEEUW, P. J. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. Journal of Computational and Applied Mathematics, v. 20, p. 53–65, 1987.

WASKOM, M. et al. **Seaborn: statistical data visualization**. Disponível em: https://seaborn.pydata.org/